Leveraging Uncertainty for Finnish L2 Speech Scoring with LLMs

Ekaterina Voskoboinik, Nhan Phan, Tamás Grósz, Mikko Kurimo
Department of Information and Communications Engineering
Aalto University, Finland
firstname.lastname@aalto.fi

Abstract

Automatic speech assessment (ASA) supports learning but often requires extensive data, which is scarce for languages with fewer learners. Recent research shows that Large Language Models (LLMs) can generalize to new tasks with minimal training data using in-context learning (ICL). We find LLMs effective in estimating the proficiency of individuals learning Finnish as a second language (L2) when given a few examples of human expert grading. The proficiency grades produced by the model, when evaluating verbatim transcripts from an automatic speech recognition (ASR) system, agree with human ratings at a level comparable to the agreement between the human raters. Our experiments reveal that adding more grading demonstrations in ICL improves the model's accuracy but, counterintuitively, increases its uncertainty when selecting an appropriate proficiency level. We show that this uncertainty can be leveraged further by creating soft labels: instead of assigning the most probable level (hard label), we aggregate the model's confidence across all possible levels, resulting in noticeable performance improvements. Further analysis reveals that the sources of model uncertainty differ across ICL settings. In zero-shot, uncertainty stems from intrinsic response properties, such as proficiency level. In few-shot, it is driven by the relationship between the sample and the demonstrations.

1 Introduction

In this study, we focus on automatically assessing the proficiency of second-language (L2) speakers producing spontaneous Finnish speech. Automatic Speech Assessment (ASA) holds significant potential for supporting language learning. However, ASA systems typically depend on machine learning algorithms, which are challenging to train when the available data is limited or when certain classes (proficiency levels in ASA case) are underrepresented. Consequently, the development of ASA systems may be hindered by the scarcity and class imbalance of annotated data. These challenges are particularly pressing for languages with smaller learner bases, such as Finnish, where data availability is inherently limited. Ironically, these resource-limited languages are likely to benefit the most from automated systems that support language learners.

Early ASA approaches for L2 data used models with hand-crafted features targeting specific aspects of spoken proficiency, such as delivery (pronunciation, fluency), language use (vocabulary, grammar), and content (Zechner et al., 2009; Bernstein et al., 2010; Chen and Zechner, 2011; Xie et al., 2012). These features were selected to align with scoring rubrics, ensuring they were meaningful and interpretable within constructs of communicative competence. Later, these hand-crafted features were replaced by representations extracted by neural networks, leading to improved model performance (Chen et al., 2018; Qian et al., 2019; Yoon and Lee, 2019).

However, both hand-crafted features and traditional neural approaches rely on large amounts of labeled data, which is often scarce for L2 ASA. Pre-trained text and audio models have shown themselves as a successful solution to this issue (Wang et al., 2021; Bannò and Matassoni, 2023). Transformer-based models (Vaswani et al., 2017), like BERT (Devlin, 2018) and wav2vec 2.0 (Baevski et al., 2020), are trained in a self-supervised manner on large unlabeled data to learn meaningful language representations. These mod-

Proceedings of AAAS Workshop 2025 – Automatic Assessment of Atypical Speech, pages 7–15
March 5, 2025 ©2025 University of Tartu Library

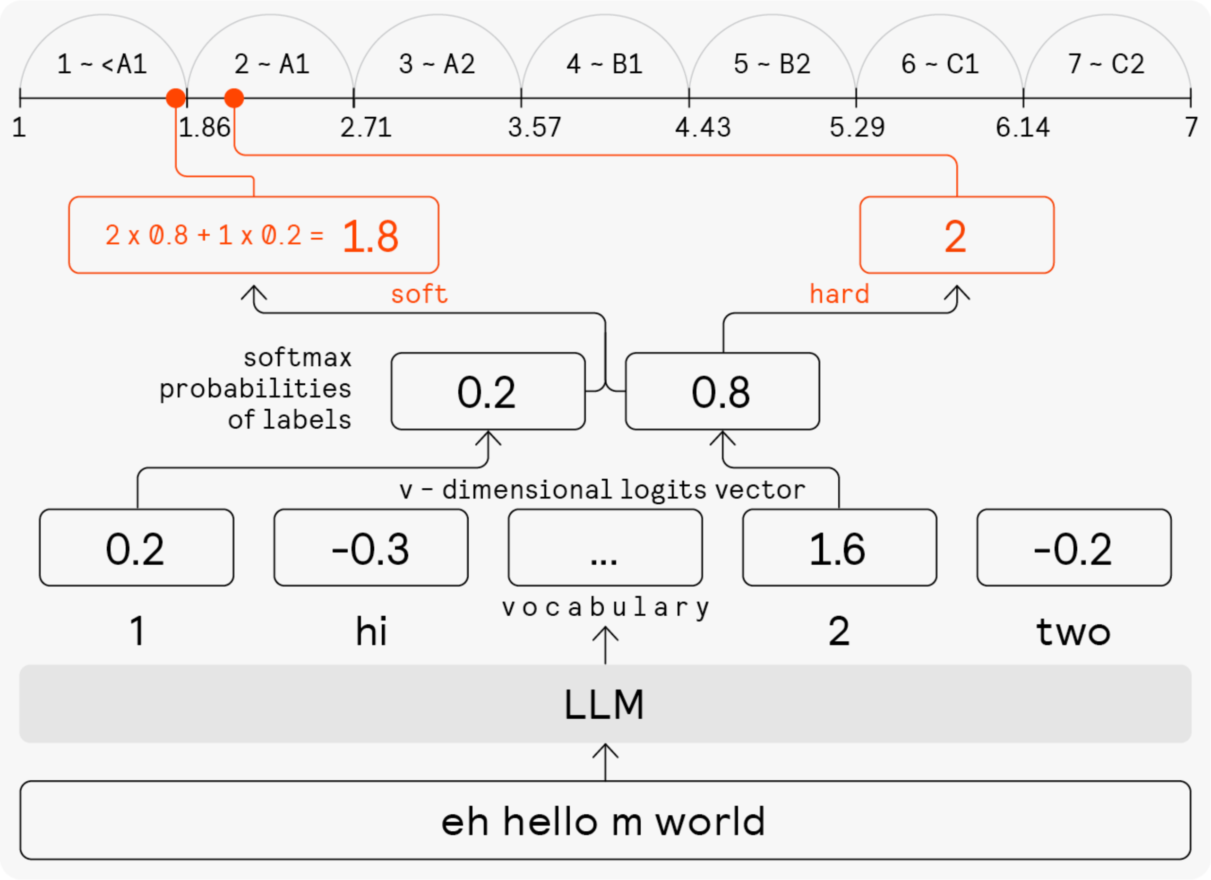

Figure 1: An illustration of assigning a speech sample into two classes through soft and hard labeling. An LLM produces a vector of logits for each token in the vocabulary. The logits corresponding to the class labels (1 and 2) are selected and transformed into a probability distribution using a softmax function. In hard labeling, the class with the highest probability (class 2 with 0.8) is selected. In soft labeling, the probabilities are used as weights: 1 is multiplied by 0.2, and 2 is multiplied by 0.8, resulting in an aggregated score of 1.8. To determine the final level, the aggregated score is mapped to its corresponding bin (bin one in this case as opposed to bin two in hard labeling).

els can then be fine-tuned on smaller datasets for specific tasks. Audio-based models have gained popularity in ASA for bypassing automatic speech recognition (ASR) by directly capturing content, language use, and delivery. They are effective not only for L2 English but also for languages with smaller learner populations, such as Finnish and Finland Swedish (Al-Ghezi et al., 2023). However, these models still face challenges with class imbalance, even when techniques like oversampling and curriculum learning are applied (Lun et al., 2024).

Recent research shows that large language models (LLMs) generalize effectively to tasks with minimal or no annotated data (Radford et al., 2019; Brown, 2020) and possess an implicit understanding of language proficiency (Malik et al., 2024; Kobayashi et al., 2024), making them a promising avenue for addressing the challenges of low-resource Finnish L2 ASA. In this study, we test whether LLMs can effectively differentiate proficiency levels in Finnish L2 speech. We examine how the model's decisions evolve across different in-context learning (ICL) settings (Brown, 2020): where the model is either prompted with the instruction of how to evaluate spoken proficiency or with instructions and grading examples. We observe that while performance improves with more examples, the model becomes less confident in its predictions, distributing probabilities more evenly across levels. To take advantage of this uncertainty, we explore soft labeling, where probabilities across all levels are aggregated as opposed to hard labeling, which assigns only the most probable level. Finally, we analyze the characteristics of responses where soft and hard labels differ, to better understand what makes the model more uncertain and how this uncertainty contributes to grading performance.

2 Data

The data used in this study is a subset of the DigiTala dataset[1], featuring speech samples from learners of Finnish. These samples include responses to semi-structured and open-ended tasks completed by university and upper secondary school students in Finland. Each response was rated on multiple dimensions: pronunciation, fluency, accuracy, range, task completion, and holistically. In this study, we focus on holistic scores as they demonstrated the highest agreement between human raters. The scores range from 1 to 7, corresponding to levels from below A1 to C2 of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). The CEFR framework evaluates spoken proficiency holistically, encompassing not only delivery features but also language use and content. For samples with multiple ratings, the scores were averaged and mapped to one of seven equal bins within the 1–7 scale to produce an integer score.

For this research, a subset of three tasks was selected based on several criteria: relatively strong human-to-human agreement compared to other tasks (as measured by quadratically weighted kappa (QWK)); task prompts designed to elicit responses of varying lengths; and representation from different student populations (school vs. university). Tasks A and B, performed by school students, involved describing their important place and a library picture, respectively, while Task C, for university students, asked them to talk about their day. When combined, the overall inter-rater agreement across these three tasks, as measured by QWK, is 0.73. Transcriptions were created by human transcribers who recorded speech verbatim, including mispronunciations and hesitations. These transcripts were used to train a wav2vec 2.0 ASR model, which was fine-tuned on native Finnish and then adapted to L2 speech, achieving word and character error rates (WER and CER) of 21.08% and 6.08%, respectively, on the entire L2 Finnish subset of the DigiTala dataset. No external language models or vocabulary were used, allowing the transcripts serve as proxies for certain delivery features, such as mispronunciations.

	A	B	C
Number of responses	173	63	106
Average duration (s)	43.32	36.00	57.27
QWK	0.50	0.61	0.41
Average score	5.16	4.17	2.80
WER (%)	18	22	35

Table 1: Task Statistics

Table 1 summarizes the statistics for each task. Notably, the QWK values indicate low agreement among human raters, as this metric accounts for the magnitude of disagreements by penalizing larger differences more heavily. Figure 2 shows the imbalanced level distributions, further highlighting the challenge of proficiency scoring for data-driven algorithms.

3 Methods

3.1 Prompting and In-context Learning

LLMs solve tasks through next-token prediction, guided by an input text or "prompt" that specifies the task or instructions. In zero-shot prompting, the model receives only instructions on how to perform a task, without examples. In contrast, in-context learning (ICL) includes demonstrations: one-shot provides a single example, and few-shot offers multiple. For ASA proficiency scoring with LLMs used in this work, an example consists of a response ASR transcript and its corresponding score from human raters.

Chat-tuned models (Touvron et al., 2023) utilize prompts designed to simulate conversational roles, marked by special tokens for system, user, and assistant turns. In this format, the system message contains grading instructions, the user message provides the response transcript, and the assistant message delivers the score. In zero-shot prompts, only system and user messages are included, whereas in ICL, user-assistant message pairs with human grading examples are also injected. The example of one-shot prompt with a chat-tuned model is given in Figure 3.

3.2 Hard vs Soft Labeling

Instead of relying on the model to generate an output score directly or selecting the most probable level token (hard labeling), we use a soft labeling approach by aggregating the model's confidence

[1] https://www.kielipankki.fi/corpora/digitala/

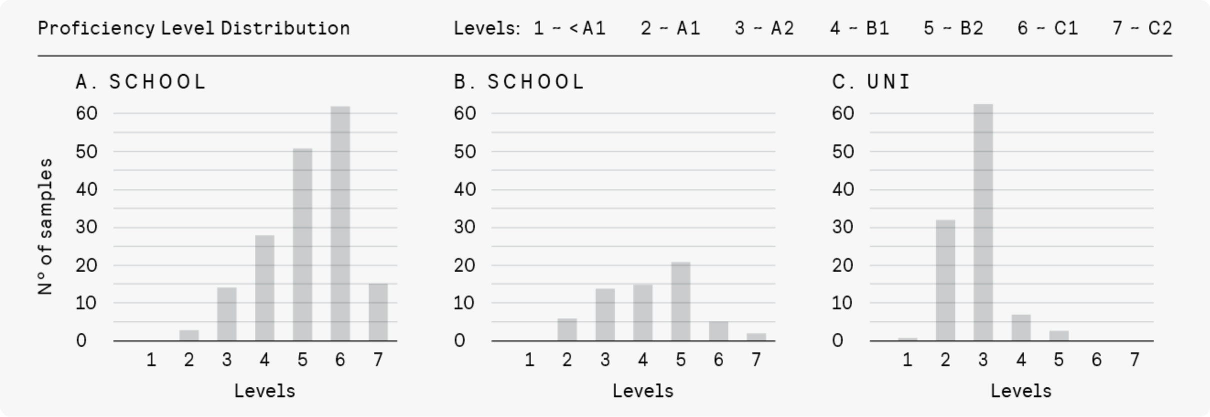

Figure 2: The distribution of responses with different proficiency levels among the tasks.

across all possible proficiency levels. We first collect logits for each level, then apply the softmax function to convert them into a probability distribution. This distribution is used to compute a weighted average label. For example, if the model assigns 80% confidence to level 2 and 20% to level 1, the weighted average is 1.8. This score is then mapped to the bins used for converting average human ratings to integers. Figure 1 illustrates the soft labeling process.

3.3 Entropy as Model's Uncertainty Measure

Entropy measures the uncertainty in a probability distribution. When probability mass is concentrated on one class, entropy is low; when all labels are equally likely, entropy is high. We compute entropy for the proficiency label space (1-7) using logits from the model's next-token prediction:

$$\text{Entropy} = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

where n is the number of labels (7), and $P(x_i)$ is the probability of label i after applying softmax. This reflects the uncertainty of the model when determining which proficiency level to assign to a student response.

3.4 Response Characteristics

Here, we describe the response properties explored in relation to their influence on model uncertainty.

Perplexity: Perplexity (ppl) measures how "surprised" a language model is by a sequence of tokens, with lower values indicating that the model can predict the next token more accurately. We calculate it as the exponentiated average negative log-likelihood of the tokens in the user message, conditioned on the previous prompt:

$$\text{Perplexity} = \exp\left(-\frac{1}{t}\sum_{i=1}^{t} \log P(x_i \mid \text{context})\right)$$

where t is the number of tokens in the test sample, x_i are the tokens, and $P(x_i \mid \text{context})$ is the conditional probability given the prompt. The context differs by prompt setting: zero-shot uses only system instructions, while ICL also includes demonstration examples before the sample transcript.

Human Variance: Most recordings received multiple ratings, with raters often disagreeing, as shown by the QWK values in Table 1. We measure this uncertainty using the variance of human scores for each response.

Demonstration Proximity: Prior research suggests that good demonstrations are often semantically close to the graded sample (Liu et al., 2021). We measure this proximity using cosine distance between embeddings generated by the LLM. In the few-shot setting, we calculate the average distance to all the demonstrations. Demonstrations are embedded with human transcripts, while test samples use ASR transcripts. Specifically, we compute embeddings for each token in the transcript using the LLM's encoder and then average these token-level embeddings to create a fixed-length representation for the entire transcript.

CEFR Level: We also use the average unbinned CEFR score to examine whether the model's uncertainty in grading is influenced by the proficiency level of the response.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a system designed to evaluate the language proficiency level of verbal
responses from students learning {language}. Your input will be a verbatim
transcript of their spoken response. Your task is to assign a proficiency level
(ranging from 1 to 7) based on the provided proficiency scale:

{proficiency_scale}

You are required to evaluate responses to the following language test task
instruction:

"{task_description}"

Your response should contain only the level, formatted as follows:
Level: X

Please adhere strictly to this format.<|eot_id|><|start_header_id|>user<|end_
header_id|>

{response_transcript_demonstration}<|eot_id|><|start_header_id|>assistant<|end_
header_id|>

Level: {response_score_demonstration}<|eot_id|><|start_header_id|>user<|end_
header_id|>

{response_transcript}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Level:
```

Figure 3: Example of a one-shot prompt in a chat-tuned LLM. Text in orange shows the tokens used by the model to differentiate between system, user, and assistant roles.

4 Experiments

4.1 Model and Prompts

Model: The LLM used in this work is Llama 3.1, an 8-billion parameter model tuned for chat.[2]
Prompts: All prompts start with a system message containing instructions to grade proficiency, the grading criteria used by raters, and the task instructions given to students. For the picture task, the picture description was included since the model is text-only. For ICL demonstrations, we selected a response from each score bin with full rater agreement or, if none were available, with minimal disagreement ($\leq$ 1-point). Demonstrations were fixed and not used as test samples. In one-shot, a random demonstration from the same task was used, while in few-shot, all demonstrations were included in a consistent random order. Each prompt ended with the assistant message formatted as “Level: ” to ensure the next token predicted was the proficiency level. The example of a one-shot prompt used in this study is shown in Figure 3. [3]

4.2 Response Characteristics Analysis

To understand what makes the model uncertain, we test whether the characteristics of samples with matching hard and soft labels differ significantly from those with different labels, using Mann-Whitney U tests across zero-, one-, and few-shot settings.

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[3] The code and prompt variables are available at https://github.com/katildakat/LLM_ASA_SOFT_LABELS.

5 Results

5.1 Proficiency Scoring and Model Uncertainty

Table 2 presents proficiency scoring results measured by accuracy (Acc), macro F1, QWK, and macro mean absolute error (MAE). Macro indicates that the metric was computed for responses in each level and then averaged to boost the influence of the underrepresented classes for the final score. The table also includes how often soft labels are closer to the true label than hard labels (denoted as “S wins”) shown as fractions (e.g., 10/30 means soft labels were closer in 10 out of 30 cases where soft and hard labels differ). It also includes the average model uncertainty, quantified by the entropy H of the probability distribution over level tokens.

The results show that ICL approaches consistently outperform zero-shot, with performance improving as the number of examples increases. Few-shot learning achieves the best results across all metrics. Notably, model uncertainty increases with the number of demonstrations (the entropy rises from 0.75 in zero-shot to 1.19 in few-shot). This growing uncertainty aligns with an increasing benefit of soft labeling: in zero-shot, hard labels outperform soft labels most of the time (12/49), but soft labeling shows a slight advantage in one-shot (23/40) and a substantial improvement in few-shot (74/117). These trends are reflected in other performance metrics, highlighting the value of soft labeling when paired with few-shot ICL.

	Acc↑	F1↑	QWK↑	MAE↓	S wins	H
z_H	.26	.15	.21	1.63		
z_S	.24	.14	.23	1.68	12/49	0.75
o_H	.24	.18	.39	1.34		
o_S	.26	.18	.43	1.29	23/40	0.86
f_H	.31	.24	.61	1.05		
f_S	**.36**	**.30**	**.67**	**0.93**	74/117	1.19

Table 2: Proficiency scoring results with hard (H) and soft (S) labeling. Metrics include accuracy (Acc), macro F1, QWK (↑ better), and macro MAE (↓ better). “S wins” shows how often soft labels outperform hard labels, and H denotes model uncertainty (entropy). z_H/S, o_H/S, and f_H/S represent zero-, one-, and few-shot learning, respectively.

Figure 4 shows the average probability distributions of proficiency label tokens across zero-, one-, and few-shot settings, illustrating how model uncertainty evolves with increasing contextual information. Each line represents the model’s predicted probabilities for a true proficiency level. In zero-shot, the distribution is narrow, with most responses concentrated around level 3, reflecting lower uncertainty and more conservative decisions. As more examples are provided, the distributions spread out, with few-shot showing the widest divergence and highest entropy. This increased uncertainty in few-shot settings enables more nuanced and less deterministic decision-making, which is also why soft labeling differs most significantly from hard labeling in this setting.

5.2 Response Characteristics Analysis

Table 3 compares the characteristics of responses where hard and soft labels match to those where they differ. This analysis aims to identify the properties of a sample that make the model less confident in predicting a single level during evaluation. The arrow direction in the table indicates whether the characteristic value increases (↑) or decreases (↓) for samples where soft labeling diverges from hard labeling.

Entropy is included as a sanity check to ensure that the model exhibits higher uncertainty for samples where soft labels differ from hard labels, and indeed, the results show that entropy is consistently higher (↑) across all ICL settings. This aligns with expectations, as higher uncertainty allows non-dominant classes to shift the probability away from the dominant label. The factors driving this uncertainty vary across settings: in zero-shot, both perplexity (↑) and CEFR score (↓) significantly influence entropy. In one-shot, none of the other tested characteristics show significant differences. In few-shot, cosine distance (↓) emerges as a key factor, indicating that responses closer to the demonstrations tend to have higher uncertainty

6 Discussion

Our results confirm previous findings (Brown, 2020) that ICL outperforms zero-shot, with performance improving as more demonstrations are added. The best setup (few-shot with soft labeling) achieves human-level agreement, reaching a QWK of 0.67 compared to 0.73 for human raters. A macro MAE of just 1 point indicates reliable differentiation between proficiency levels. While

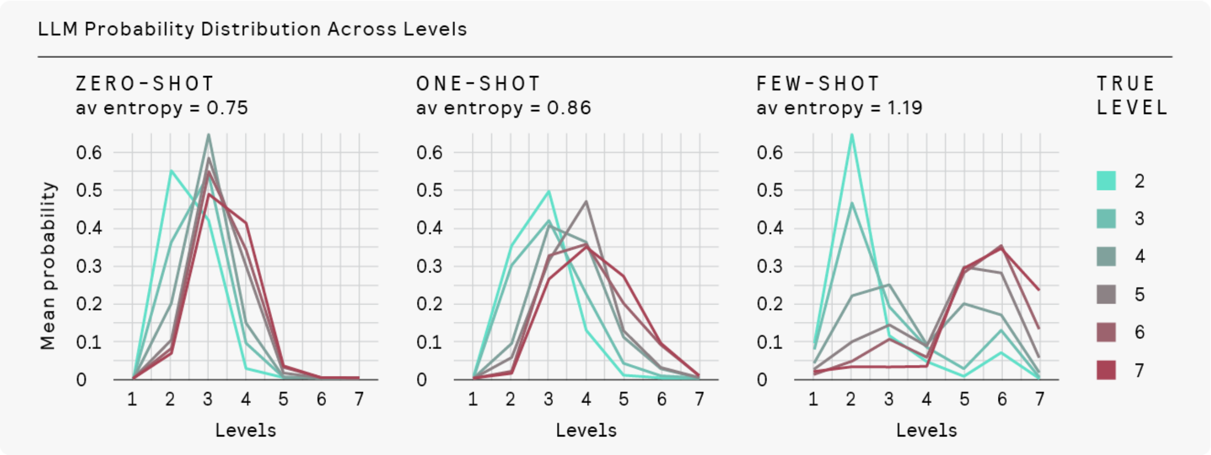

Figure 4: Average probability distributions for zero-, one-, and few-shot settings.

	zero	one	few
entropy	✓↑	✓↑	✓↑
ppl	✓↑	✗	✗
human variance	✗	✗	✗
cosine distance	–	✗	✓↓
CEFR level	✓↓	✗	✗

Table 3: Comparison of response characteristics for samples where hard and soft labels match versus those where they differ. Rows represent characteristics and columns represent ICL settings (zero-, one-, few-shot). Arrows indicate whether the characteristic increases (↑) or decreases (↓) for samples with differing soft and hard labels.

accuracy and macro F1 remain modest, this reflects the data's challenging nature, even for human raters.

In ICL, entropy increases with more demonstrations, yet this added uncertainty enhances performance, particularly with soft labeling. We suspect that higher entropy indicates the model's learning of proficiency level cues. Interestingly, in few-shot settings, entropy is higher when demonstrations are closer to the test sample, even though closer examples do improve predictions (Liu et al., 2021), one would expect this to occur with less uncertainty, not more.

Consistent with (Sánchez et al., 2024), we find a negative correlation between perplexity and CEFR levels (-0.59 Spearman r) and a positive correlation between perplexity and WER (0.75 Spearman r), suggesting that beginner learners tend to produce speech that deviates more from what LLMs and ASR models consider well-formed. However, high perplexity (and thus WER) only affects predictions in zero-shot, where the model becomes uncertain and acts as a severe rater according to its bias. In ICL, perplexity does not influence uncertainty, as the model relies on the relationship between the sample and the demonstrations to base its decisions on.

Surprisingly, human disagreement did not affect the LLM's decisions. This could be due to the study's limitations: the models only had access to ASR transcripts, which serve as proxies for pronunciation and fluency. Delivery features that may contribute to human score variance were not available.

7 Conclusion

This study shows that LLMs can effectively grade Finnish L2 speech in a few-shot setting, achieving QWK scores comparable to human raters, with soft labeling being especially beneficial. More demonstrations in ICL increase entropy, enhancing performance in few-shot prompting. In ICL, perplexity does not influence scoring decisions; rather, the model's uncertainty is shaped by the relationship between the sample and demonstrations, suggesting that closer proximity in the embedding space helps the model identify nuanced cues. We think that soft labeling can be valuable not only for language proficiency scoring but also for other ordinal classification tasks. Future work will focus on fine-tuning the model to include Finland Swedish and incorporating delivery features directly to LLMs to improve grading accuracy.

Acknowledgements

This work was funded by Research Council of Finland's grants 322625, 345790, 355587 and 365233. The computational resources were provided by Aalto ScienceIT.

References

Ragheb Al-Ghezi, Yaroslav Getman, Ekaterina Voskoboinik, Mittul Singh, and Mikko Kurimo. 2023. Automatic rating of spontaneous speech for low-resource languages. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 339–345. IEEE.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Stefano Bannò and Marco Matassoni. 2023. Proficiency assessment of l2 spoken english using wav2vec 2.0. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1088–1095. IEEE.

Jared Bernstein, Jian Cheng, and Masanori Suzuki. 2010. Fluency and structural complexity as predictors of l2 oral proficiency. In *Eleventh annual conference of the international speech communication association*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. 2018. End-to-end neural network based automated speech scoring. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6234–6238. IEEE.

Miao Chen and Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 722–731.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Large language models are state-of-the-art evaluator for grammatical error correction. *arXiv preprint arXiv:2403.17540*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Tin Lun, Ekaterina Voskoboinik, Ragheb Al-Ghezi, Tamás Grósz, and Mikko Kurimo. 2024. https://doi.org/10.21437/Interspeech.2024-760 Oversampling, augmentation and curriculum learning for speaking assessment with limited training data. pages 4019–4023.

Ali Malik, Stephen Mayhew, Chris Piech, and Klinton Bicknell. 2024. From tarzan to tolkien: Controlling the language proficiency level of llms for content generation. *arXiv preprint arXiv:2406.03030*.

Yao Qian, Patrick Lange, Keelan Evanini, Robert Pugh, Rutuja Ubale, Matthew Mulholland, and Xinhao Wang. 2019. Neural approaches to automated speech scoring of monologue and dialogue responses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 8112–8116. IEEE.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ricardo Muñoz Sánchez, Simon Dobnik, and Elena Volodina. 2024. Harnessing gpt to study second language learner essays: Can we use perplexity to determine linguistic competence? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 414–427.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Xinhao Wang, Keelan Evanini, Yao Qian, and Matthew Mulholland. 2021. Automated scoring of spontaneous speech from young learners of english using transformers. In *2021 IEEE spoken language technology workshop (SLT)*, pages 705–712. IEEE.

Shasha Xie, Keelan Evanini, and Klaus Zechner. 2012. Exploring content features for automated speech scoring. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pages 103–111.

Su-Youn Yoon and Chungmin Lee. 2019. Content modeling for automated oral proficiency scoring system. In *Proceedings of the fourteenth workshop* *on innovative use of NLP for building educational applications*, pages 394–401.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. https://api.semanticscholar.org/CorpusID:27619107 Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Commun.*, 51:883–895.